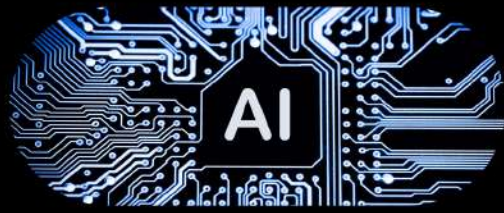


AI ENGINEERING SKILLS CHECKLIST



TECHNICAL FOUNDATION

- ☐ Basic statistics and probability
- ☐ Basic linear algebra
- ☐ Basic calculus
- ☐ Understanding of numerical precision formats
- ☐ Data structures and algorithms fundamentals

SOFTWARE ENGINEERING

- ☐ Python programming
- ☐ Linux & command-line tools
- ☐ Git and version control
- ☐ API design and integration
- ☐ Basic understanding of Docker and containers
- ☐ Cloud platforms
- ☐ System architecture concepts
- ☐ Monitoring and logging
- ☐ Testing

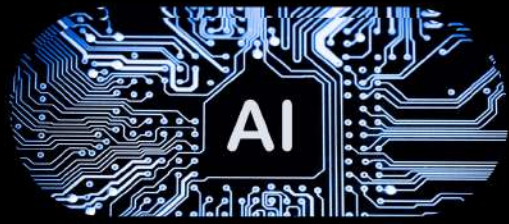
ML BASICS

- ☐ Understanding of supervised/unsupervised learning
- ☐ Common algorithms at a high level
- ☐ Model evaluation metrics
- ☐ Training/validation/test splits
- ☐ Overfitting and underfitting
- ☐ Strong deep learning knowledge

FOUNDATION MODELS & MODEL SELECTION

- ☐ Transformer architecture
- ☐ Attention mechanism
- ☐ Tokenization
- ☐ Model scaling laws (e.g. Chinchilla)
- ☐ Post-training techniques: SFT, RLHF, DPO
- ☐ Tradeoffs: cost vs performance vs licensing
- ☐ Open-weight vs open-source vs API models
- ☐ Tooling for model benchmarking

AI ENGINEERING SKILLS CHECKLIST



EVALUATION AND TESTING

- ☐ Model evaluation pipelines
- ☐ Evaluation metrics: perplexity, BLEU, ROUGE, semantic similarity, functional correctness, etc.
- ☐ Using AI judges and human evals
- ☐ Measuring hallucinations, toxicity, bias

PROMPT ENGINEERING

- ☐ Structuring effective prompts
- ☐ In-context learning techniques
- ☐ Defensive prompt engineering against attacks
- ☐ Prompt experimentation and tracking

RETRIEVAL-AUGMENTED GENERATION (RAG)

- ☐ Vector database implementation
- ☐ Document chunking strategies
- ☐ Embedding techniques
- ☐ Term-based vs. embedding-based retrieval
- ☐ Retrieval optimization techniques

AGENT SYSTEMS

- ☐ Tool integration
- ☐ Planning techniques
- ☐ Memory systems implementation
- ☐ Agent security and safety guardrails
- ☐ Agent evaluation methodologies

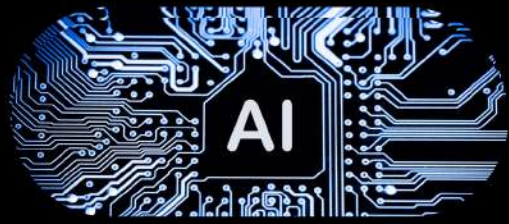
FINETUNING

- ☐ Parameter-efficient fine-tuning (PEFT)
- ☐ LoRA and similar approaches
- ☐ Model distillation
- ☐ Model merging
- ☐ Multi-task fine-tuning

DATASET ENGINEERING

- ☐ Data acquisition strategies
- ☐ Data quality assessment
- ☐ Data processing
- ☐ Annotation guidelines creation
- ☐ Data augmentation and synthesis

AI ENGINEERING SKILLS CHECKLIST



INFERENCE OPTIMIZATION

- ☐ Understanding compute vs memory-bound inference
- ☐ Latency metrics: TTFT, TPOT
- ☐ Model compression: quantization, pruning, distillation
- ☐ Batch vs. online inference strategies
- ☐ Hardware (GPU, TPU, memory specs)
- ☐ Batching techniques
- ☐ Parallel inference strategies
- ☐ Caching implementations

USER FEEDBACK INTEGRATION

- ☐ Feedback system design
- ☐ Explicit vs. implicit feedback collection
- ☐ Data "flywheels"
- ☐ Human-in-the-loop approaches
- ☐ Continuous improvement cycles

APPLICATION ARCHITECTURE

- ☐ Context construction patterns
- ☐ Input/output guardrails
- ☐ Model routing and gateways
- ☐ Caching architectures
- ☐ Orchestration patterns

SECURITY/PRIVACY/ETHICS

- ☐ Prompt injection detection/mitigation
- ☐ Adversarial input handling
- ☐ PII detection and redaction
- ☐ Secure sandboxing (for agents/code execution)
- ☐ Model privacy risks (e.g. memorization attacks, data leakage)
- ☐ Legal compliance (GDPR, copyright implications)
- ☐ AI Ethics considerations